Performance of Maxwell GPUs and Optimization of Non-Perturbative Renormalization codes

Jeonghwan Pak, Weonjong Lee, Hwancheol Jeong, Sangbeak Lee (SNU), Jangho Kim (KISTI)

SWME collaboration

July, 14th ∼ July, 18th. 2015

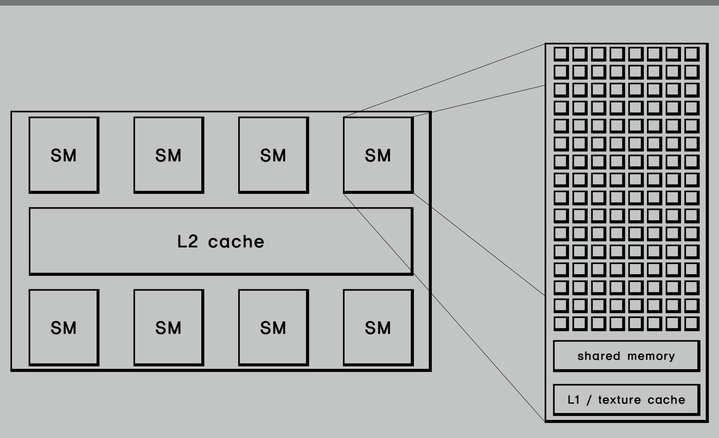
1. GTX TITAN X

GTX TITAN X

- GTX Titan X is NVIDIA's new consumer GPUs in the Maxwell generation. Titan X has good single precision(SP) performance than GPUs of Kepler generation. However, Titan X has poorer double precision(DP) calculation than GTX Titan Black and Tesla K40.

	Fermi	Kepler		Maxwell
	GTX 580	TITAN BLACK	K40	TITAN X
SP TFLOPs	1.58	5.00	4.29	6.00
DP TFLOPs	0.20	1.30	1.43	0.19
Memory Size (GB)	1.5	6	12	12
Memory Bandwidth (GB/sec)	192.4	366	288	336

SMM of Maxwell GPU

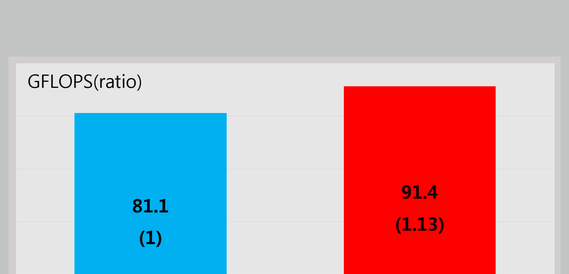


- Kepler has 64KB = shared memory + L1 cache : 16 + 48, 48 + 16, or 32 + 32.
- Maxwell has 96KB shared memory and 48KB L1 cache.

2. CG performance

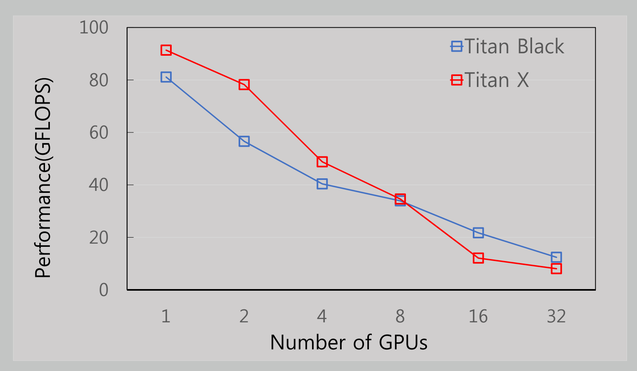
CG performance

- Conjugate Gradiant(CG) inverter code uses mixed precision algorithm, in which SP calculation is dominant.
- We measure the FLOPS of the CG using 10 configurations of $\mathbf{20^3 \times 64}$ ensembles on a single GPU.
- The performances of Titan Black & Titan X.



- CPU : i7-5930K Processor (15M Cache, up to 3.70 GHz)
- RAM : DDR4 32G PC4-17000

CG scalability



# of nodes	Titan Black	Titan X	ratio
1	81.11(0.95)	91.35(5.95)	1:1.13
2(with comunication)	56.61(0.19)	78.23(4.78)	1:1.38
2(without comunication)	55.86(0.57)	72.58(3.53)	1:1.30
4	40.38(0.44)	48.80(1.76)	1:1.21
8	33.92(0.55)	34.61(0.07)	1:1.02
16	21.74(0.55)	12.10(1.04)	1:0.56
32	12.41(1.06)	8.07(0.73)	1:0.65

3. NPR performance

NPR calculations

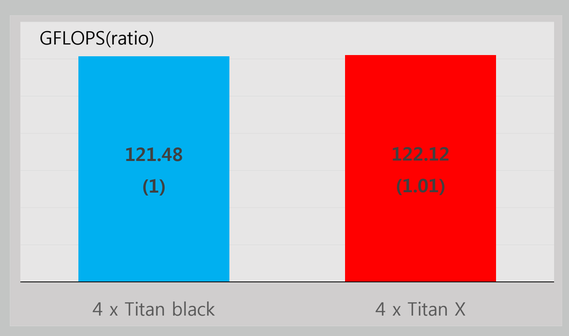
- The Non-Perturbative Renormalization(NPR) code tests the DP performance of Titan Black & Titan X.
- Dominant part is one-color four fermion operator calculation.

$$O^{f_1 f_2 f_3 f_4}_{i;I}(z) = \overline{\chi}^{f_1}_{i;c_1}(z_A)\overline{(\gamma_{S_1} \otimes \xi_{F_1})}_{AB}\chi^{f_2}_{i;c_2}(z_B) \times \overline{\chi}^{f_3}_{i;c_3}(z_C)\overline{(\gamma_{S_2} \otimes \xi_{F_2})}_{CD}\chi^{f_4}_{i;c_4}(z_D) \times [U_{i;AD}]_{c_1 c_4}(z)[U_{i;CB}]_{c_3 c_2}(z)$$

  - c1, c2, c3, c4 : $0 \sim 3$
  - S1 = S2 : $0 \sim 15$
  - A, B, C, D : $0 \sim 15$

NPR performance with old CPS

- We use 10 configurations of $\mathbf{20^3 \times 64}$ ensemble with the CPS version is 4.1.2.

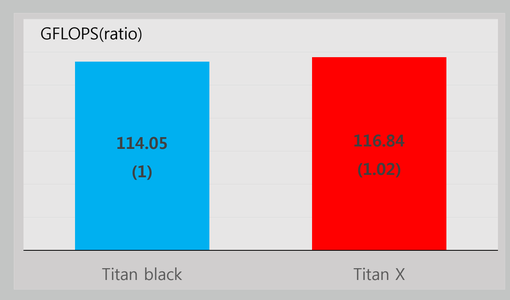


CGMA of NPR calculation

- Compute to Global Memory Access(CGMA) : the number of floating point calculations for each access to the global memory.
- (Max DP FPLOPS) = (Mem Bandwidth) × (CGMA) / (Size of double)
- CGMA of NPR code is 2.96. Hence the expected DP performance of NPR code is 124 GFLOPS on Titan Black & Titan X.

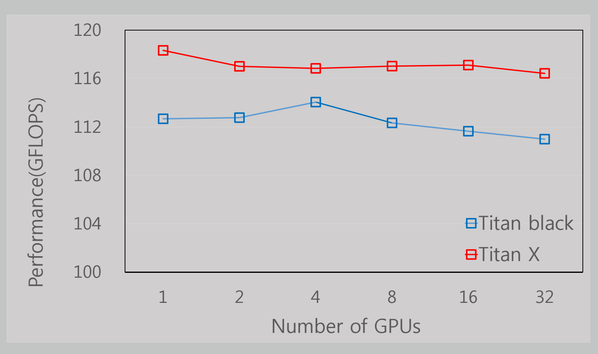
NPR performance with new CPS

- We use our CUDA code with the new CPS is V5.0.26 version included SciDAC packages (QIO,QMP), but do not use QUDA.



- NVIDIA GTX TITAN X show the top performance with a theoretical SP performance. And the theoretcal DP performance is ∼1/7 than Titan & Titan Black of the Fermi geneation. However, NPR code which is donimant DP calculation codes have similar performance without tunning than Titan & Titan Black.

NPR scalability on GPU



# of nodes	Titan Black	Titan X	ratio
1	112.67(0.67)	118.32(1.68)	1:1.05
2(with comunication)	112.77(0.85)	117.00(0.62)	1:1.04
2(without comunication)	112.23(0.79)	117.04(0.57)	1:1.04
4	114.05(1.61)	116.84(0.65)	1:1.02
8	112.33(1.30)	117.02(0.42)	1:1.04
16	111.64(0.69)	117.11(0.37)	1:1.05
32	110.99(0.09)	116.42(0.33)	1:1.05

- There is only one global sum in NPR code. In NPR code, a communication overhead is negligibly small. Hence its scalability is almost flat.

4. Conclusion

Conclusion

- Titan X in Maxwell is currently the fastest GPU in the SP calculation. The performance of the CG code, which the SP calculation is dominant, increases by 13% on Titan X compared with Titan Black.
- The DP/SP ratio of Titan X is reduced to 1/32 compared with Titan & Titan Black. However, the DP performance of Titan X is better than that of Titan Black.
- The DP performance of the NPR code on Titan X is close to its peak performance. But the DP performance of the NPR code on Titan Black is only 1/10 of its peak performance. The reason is that the bottle neck is data transfer between memories. Hence, we expect to get a better performance, if we optimize more about global memory access. In other words, we need to increase the CGMA of the NPR code.